

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problem Mailbox.**





## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<b>(51) International Patent Classification <sup>6</sup> :</b> C12Q 1/68, G06F 17/30, 159/00	<b>A1</b>	<b>(11) International Publication Number:</b> WO 99/05323 <b>(43) International Publication Date:</b> 4 February 1999 (04.02.99)
<b>(21) International Application Number:</b> PCT/US98/15151 <b>(22) International Filing Date:</b> 24 July 1998 (24.07.98)  <b>(30) Priority Data:</b> 60/053,842 25 July 1997 (25.07.97) US 60/069,198 11 December 1997 (11.12.97) US 60/069,436 11 December 1997 (11.12.97) US  <b>(71) Applicant:</b> AFFYMETRIX, INC. [US/US]; 3380 Central Expressway, Santa Clara, CA 95051 (US). <b>(72) Inventor:</b> BALABAN, David, J.; 37 Bret Harte Road, San Rafael, CA 94901 (US). <b>(74) Agents:</b> LANG, Dan, H. et al.; Townsend and Townsend and Crew LLP, 8th floor, Two Embarcadero Center, San Francisco, CA 94111-3834 (US).		<b>(81) Designated States:</b> JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).  <b>Published</b> <i>With international search report.</i>
<b>(54) Title:</b> GENE EXPRESSION AND EVALUATION SYSTEM  <b>(57) Abstract</b>  An efficient and easy to use query system for a gene expression database. Using such a system, one can easily identify genes or expressed sequence tags whose expression correlates to particular tissue types. Various tissue types may correspond to different diseases, states of disease progression, different organs, different species, etc. Researchers may now use large scale gene expression databases to full advantage.		

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

## GENE EXPRESSION AND EVALUATION SYSTEM

### CROSS-REFERENCE TO RELATED APPLICATIONS

The present application claims priority from U.S. Prov. App. No. 60/053,842  
5 filed July 25, 1997, entitled COMPREHENSIVE BIO-INFORMATICS DATABASE, from  
U.S. Prov. App. No. 60/069,198 filed on December 11, 1997, entitled COMPREHENSIVE  
DATABASE FOR BIOINFORMATICS, and from U.S. Prov. App. No. 60/069,436, entitled  
GENE EXPRESSION AND EVALUATION SYSTEM, filed on December 11, 1997. The  
contents of all three provisional applications are herein incorporated by reference.

10 The subject matter of the present application is related to the subject matter of  
the following three co-assigned applications filed on the same day as the present application:  
METHOD AND APPARATUS FOR PROVIDING A BIOINFORMATICS DATABASE  
(Attorney Docket No. 018547-033810), METHOD AND SYSTEM FOR PROVIDING A  
POLYMORPHISM DATABASE (Attorney Docket No. 018547-033820), METHOD AND  
15 SYSTEM FOR PROVIDING A PROBE ARRAY CHIP DESIGN DATABASE (Attorney  
Docket No. 018547-033830). The contents of these three applications are herein incorporated  
by reference.

### BACKGROUND OF THE INVENTION

20 The present invention relates to computer systems and more particularly to  
computer systems for analyzing expression levels or concentrations.

Devices and computer systems have been developed for collecting information  
about gene expression or expressed sequence tag (EST) expression in large numbers of tissue  
samples. For example, PCT application WO92/10588, incorporated herein by reference for  
25 all purposes, describes techniques for sequencing or sequence checking nucleic acids and  
other materials. Probes for performing these operations may be formed in arrays according to  
the methods of, for example, the pioneering techniques disclosed in U.S. Patent  
No. 5,143,854 and U.S. Patent No. 5,571,639, both incorporated herein by reference for all  
purposes.

30 According to one aspect of the techniques described therein, an array of  
nucleic acid probes is fabricated at known locations on a chip or substrate. A fluorescently

labeled nucleic acid is then brought into contact with the chip and a scanner generates an image file indicating the locations where the labeled nucleic acids bound to the chip. Based upon the identities of the probes at these locations, it becomes possible to extract information such as the monomer sequence of DNA or RNA.

5 Computer-aided techniques for monitoring gene expression using such arrays of probes have been developed as disclosed in EP Pub. No. 0848067 and PCT publication No. WO 97/10365, the contents of which are herein incorporated by reference. Many disease states are characterized by differences in the expression levels of various genes either through changes in the copy number of the genetic DNA or through changes in levels of transcription  
10 (*e.g.*, through control of initiation, provision of RNA precursors, RNA processing, etc.) of particular genes. For example, losses and gains of genetic material play an important role in malignant transformation and progression. Furthermore, changes in the expression (transcription) levels of particular genes (*e.g.*, oncogenes or tumor suppressors), serve as signposts for the presence and progression of various cancers.

15 Information on expression of genes or expressed sequence tags may be collected on a large scale in many ways, including the probe array techniques described above. One of the objectives in collecting this information is the identification of genes or ESTs whose expression is of particular importance. Researchers wish to answer questions such as: 1) Which genes are expressed in cells of a malignant tumor but not expressed in  
20 either healthy tissue or tissue treated according to a particular regime? 2) Which genes or ESTs are expressed in particular organs but not in others? 3) Which genes or ESTs are expressed in particular species but not in others?.

Collecting vast amounts of expression data from large numbers of samples including all the tissue types mentioned above is but the first step in answering these  
25 questions. To derive full value from the investment made in collecting and storing expression data, one must be able to efficiently mine the data to find items of particular relevance. What is needed is an efficient and easy to use query system for a gene expression database.

## SUMMARY OF THE INVENTION

30 An efficient and easy to use query system for a gene expression database is provided by virtue of the present invention. Using such a system, one can easily identify

genes or expressed sequence tags whose expression correlates to particular tissue types. Various tissue types may correspond to different diseases, states of disease progression, different organs, different species, etc. Researchers may now use large scale gene expression databases to full advantage.

5           According to a first aspect of the present invention, a method is provided in a computer system for operating a database storing information about compound concentration. The method includes: providing a database including concentrations of a plurality of compounds as measured in a plurality of samples, accepting a user query to the database to identify desired ones of the plurality of compounds, the user query specifying concentration  
10 characteristics of the desired compounds in selected ones of the plurality of samples, and comparing the concentration characteristics to the concentrations stored in the database to identify the desired compounds.

          A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached  
15 drawings.

### BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates an example of a computer system that may be used to execute software embodiments of the present invention.

20           Fig. 2 shows a system block diagram of a typical computer system.

Fig. 3 is a flowchart describing steps of developing expression data according to one embodiment of the present invention.

Fig. 4 is a flowchart describing steps of querying an expression database according to one embodiment of the present invention.

25           Figs. 5A-5L depict a user interface for querying an expression database according to one embodiment of the present invention.

### DESCRIPTION OF SPECIFIC EMBODIMENTS

Fig. 1 illustrates an example of a computer system that may be used to execute  
30 software embodiments of the present invention. Fig. 1 shows a computer system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have

one or more buttons such as mouse buttons 13. Cabinet 7 houses a CD-ROM drive 15 and a hard drive (not shown) that may be utilized to store and retrieve software programs including computer code incorporating the present invention. Although a CD-ROM 17 is shown as the computer readable medium, other computer readable media including floppy disks, DRAM, hard drives, flash memory, tape, and the like may be utilized. Cabinet 7 also houses familiar computer components (not shown) such as a processor, memory, and the like.

Fig. 2 shows a system block diagram of computer system 1 used to execute software embodiments of the present invention. As in Fig. 1, computer system 1 includes monitor 3 and keyboard 9. Computer system 1 further includes subsystems such as a central processor 50, system memory 52, I/O controller 54, display adapter 56, removable disk 58, fixed disk 60, network interface 62, and speaker 64. Removable disk 58 is representative of removable computer readable media like floppies, tape, CD-ROM, removable hard drive, flash memory, and the like. Fixed disk 60 is representative of an internal hard drive or the like. Other computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 50 (*i.e.*, a multi-processor system) or memory cache.

Arrows such as 66 represent the system bus architecture of computer system 1. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, display adapter 56 may be connected to central processor 50 through a local bus or the system may include a memory cache. Computer system 1 shown in Fig. 2 is but an example of a computer system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will be readily apparent to one of ordinary skill in the art. In one embodiment, the computer system is an IBM compatible personal computer.

The VLSIPS™ and GeneChip™ technologies provide methods of making and using very large arrays of polymers, such as nucleic acids, on very small chips. See U.S. Patent No. 5,143,854 and PCT Patent Publication Nos. WO 90/15070 and 92/10092, each of which is hereby incorporated by reference for all purposes. Nucleic acid probes on the chip are used to detect complementary nucleic acid sequences in a sample nucleic acid of interest (the "target" nucleic acid).

It should be understood that the probes need not be nucleic acid probes but



may also be other polymers such as peptides. Peptide probes may be used to detect the concentration of peptides, polypeptides, or polymers in a sample. The probes should be carefully selected to have bonding affinity to the compound whose concentration they are to be used to measure.

5 In one embodiment, the present invention provides methods of reviewing and analyzing information relating to the concentration of compounds in a sample as measured by monitoring affinity of the compounds to polymers such as polymer probes. In a particular application, the concentration information is generated by analysis of hybridization intensity files for a chip containing hybridized nucleic acid probes. The hybridization of a nucleic acid  
10 sample to certain probes may represent the expression level of one more genes or expressed sequence tags (EST). The expression level of a gene or EST is herein understood to be the concentration within a sample of mRNA or protein that would result from the transcription of the gene or EST.

Expression level information that is reviewed and/or analyzed by virtue of the  
15 present invention need not be obtained from probes but may originate from any source. If the expression information is collected from a probe array, the probe array need not meet any particular criteria for size and density. Furthermore, the present invention is not limited to reviewing and/or analyzing fluorescent measurements of bondings such as hybridizations but may be readily utilized for reviewing and/or analyzing other measurements.

20 Concentration of compounds other than nucleic acids may be reviewed and/or analyzed according to one embodiment of the present invention. For example, a probe array may include peptide probes which may be exposed to protein samples, polypeptide samples, or peptide samples which may or may not bond to the peptide probes. By appropriate selection of the peptide probes, one may detect the presence or absence of particular proteins,  
25 polypeptides, or peptides which would bond to the peptide probes.

A system that designs a chip mask, synthesizes the probes on the chip, labels nucleic acids from a target sample, and scans the hybridized probes is set forth in U.S. Patent No. 5,571,639 which is hereby incorporated by reference for all purposes. However, the present invention may be used separately for reviewing and/or analyzing the results of other  
30 systems for generating expression information, or for reviewing and/or analyzing concentrations of polymers other than nucleic acids.

The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular target sequence. The test probe is typically perfectly complementary to a portion (subsequence) of the target sequence. The term "mismatch control" or "mismatch probe" refer to probes whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. For each mismatch (MM) control in an array there typically exists a corresponding perfect match (PM) probe that is perfectly complementary to the same particular target sequence.

Among the important pieces of information obtained from the chips are the relative fluorescent intensities obtained from the perfect match probes and mismatch probes.

These intensity levels are used to estimate an expression level for a gene or EST. The computer system used for analysis will preferably have available other details of the experiment including possibly the gene name, gene sequence, probe sequences, probe locations on the substrate, and the like.

An expression analysis is performed for each gene for each experiment. Fig. 3 is a flowchart describing steps of estimating an expression level for a particular gene as measured in a particular experiment on a chip. At step 302, the computer system receives raw scan data of N pairs of perfect match and mismatch probes. In a preferred embodiment, the hybridization intensities are photon counts from a fluorescein labeled target that has hybridized to the probes on the substrate. For simplicity, the hybridization intensity of a perfect match probe will be designed " $I_{pm}$ " and the hybridization intensity of a mismatch probe will be designed " $I_{mm}$ ".

Hybridization intensities for a pair of probes are retrieved at step 304. The background signal intensity is subtracted from each of the hybridization intensities of the pair at step 306. Background subtraction can also be performed on all the raw scan data at the same time.

At step 308, the hybridization intensities of the pair of probes are compared to a difference threshold (D) and a ratio threshold (R). It is determined if the difference between the hybridization intensities of the pair ( $I_{pm} - I_{mm}$ ) is greater than or equal to the difference threshold AND the quotient of the hybridization intensities of the pair ( $I_{pm} / I_{mm}$ ) is greater than or equal to the ratio threshold. The difference thresholds are typically user defined values that have been determined to produce accurate expression monitoring of a gene or

genes. In one embodiment, the difference threshold is 20 and the ratio threshold is 1.2.

If  $I_{pm} - I_{mm} \geq D$  and  $I_{pm} / I_{mm} \geq R$ , the value NPOS is incremented at step 310. In general, NPOS is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely expressed. NPOS is utilized in a determination of the expression of the gene.

At step 312, it is determined if  $I_{mm} - I_{pm} \geq D$  and  $I_{mm} / I_{pm} \geq R$ . If these expressions are true, the value NNEG is incremented at step 314. In general, NNEG is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely not expressed. NNEG, like NPOS, is utilized in a determination of the expression of the gene.

For each pair that exhibits hybridization intensities either indicating the gene is expressed or not expressed, a log ratio value (LR) and intensity difference value (IDIF) are calculated at step 316. LR is calculated by the log of the quotient of the hybridization intensities of the pair ( $I_{pm} / I_{mm}$ ). The IDIF is calculated by the difference between the hybridization intensities of the pair ( $I_{pm} - I_{mm}$ ). If there is a next pair of hybridization intensities at step 318, they are retrieved at step 304.

For each analysis performed certain data is stored in an expression analysis database. There is preferably a record for each gene or EST for which the chip measures expression. This record includes fields to hold various pieces of information. One field stores an analysis ID to identify the analysis. A result type ID field indicates whether the listed expression results indicate that the gene is present, marginal, absent, or unknown based on application of a decision matrix to the values P1, P2, P3, and P4. A number\_positive field shows NPOS. A number\_negative field shows NNEG. A number\_used field shows the number of probes belonging to pairs that incremented NNEG or NPOS. A number\_all field indicates N. An average log ratio field indicates the average LR for all probe pairs. A number\_positive\_exceeds field indicates the value of NPOS - NNEG. A number\_negative\_exceeds field indicates the value of NNEG - NPOS. An average differential intensity field indicates the average IDIF for the probe pairs. A number\_in\_average field indicates the number of probe pairs used in computing the average.

Steps of operating a user interface to the expression database will now be illustrated with reference to Fig. 4. The steps of Fig. 4 may be repeated or may occur in a

different order, or one or more steps may be omitted. The discussion of the user interface will also refer to Figs. 5A-5L which depict representative screen displays of the user interface.

At step 402, the user selects files of expression analysis results for querying.

- 5 Fig. 5A illustrates an interface screen where the user may specify expression results files. Each file represents one experiment. A table 502 lists the files that have already been selected. A given list may be saved for later use by selecting a button 504. A previously saved list may be deleted by selecting a button 506. A button 508 resets the list depicted in table 502 to a previously saved version. An import button 512 imports the contents of the
- 10 files depicted in table 502 for querying. Within table 502, a file name column lists the file names that would be imported by application of import button 512. A code column indicates the tissue type for the expression data in each file. A replicate file indicates whether the file is a duplicate. A chip design code column indicates the chip design used to generate the data for the file. Various other columns (not shown) give further information about the analysis
- 15 result data.

- By selecting a select files button 514, the user calls up a select files screen 516 as shown in Fig. 5B. This provides an interactive file search and selection process that does not require typing in the file name. Before importing the file list, the user should select a species by using a species drop-down list 518 as shown in Fig. 5C. An analysis-type drop
- 20 down list 519 allows the user to select between a relative expression analysis and an absolute expression analysis.

- Fig. 5D shows a normalization form 520 for normalizing imported expression results at step 404. The software scales the average difference data generated by the analysis routine based on the user's selections on normalization form 520. In a chip variability area
- 25 522, the user specifies housekeeping genes with known expression levels and selects a scale value. The user can elect to either apply or not apply this scale value. If the user elects to apply the scale value, each gene expression level measured on a single chip is multiplied by a value equal to the desired scaling factor divided by the average of housekeeping expression levels measured on that chip.

- 30 Also on normalization form 520, in a tissue variability area 524, the user may select a scale value that applies to data collected from multiple chips and whether or not it is

applied. If this scale value is to be applied, each expression value measured in a chip set is multiplied by a factor equal to the scale value divided by the average expression level measured over all genes for the entire chip set. A transformation area 526 allows the user to select whether negative average difference values are to be converted to positive numbers by use of a logarithmic transform. The user can reset all the changes made on normalization form 520 by selecting a reset button 528 or apply the selected normalizations and transformations by selecting an apply button 530.

At step 406, the user filters the large set of experimental data that was imported, normalized, and transformed. Fig. 5E depicts a filter experiments form 532. A lower table 534 lists the imported experiments and genes or EST and the expression data associated with each combination of experiment and gene or EST. An upper table 536 is used to enter a query to filter the experiment data in lower table 534. Each column of upper table 536 corresponds to a column in lower column 534. Upper table 536 is similar to a query by example (QBE) grid as included in Microsoft Access. Predicates are entered in the columns of upper table 536 with all the predicates in a single row treated as ANDs and those between rows treated as OR's. The results satisfying a given query are displayed in lower table 534 upon selection of a filter button 538. Filters may be saved, deleted, and reset by use of appropriately labeled buttons, 540, 542, and 544. A stored filter may be loaded by use of a drop-down list 546. Selection of an export button 547 writes the data to an Excel spreadsheet.

20

To facilitate further user queries, the user may specify a new field to be used as a pivot field for future queries at step 408. Elements of the selected field will become columns in the new table. Fig. 5F shows how a pivot value is selected by use of a drop-down list 548. The pivot value identifies the expression data that will be listed in the columns of lower table 534. Fig. 5G shows a pivot column drop-down list 550 allows selection of a particular column of lower table 534 as the pivot field. The entries of the selected column are shown in a left list box 552 and moved to a right list box 554 to include them as rows in the pivoted table. The user selects arrow keys 556 to add and delete items of right list box 554. To perform the pivot operation, the user selects a pivot button 558.

30

Fig. 5H depicts a user interface for filtering tissue types as displayed as a result of the pivot operation. Lower table 534 shows the result of a pivot operation as

described with reference to Figs. 5F-5G.

Upper table 536 is now used at step 410 to specify a query to filter genes using the results of experiments obtained from different tissue types. Again, predicates in a row are treated as ANDs. Predicates between rows are treated as ORs. By properly formulating a query, the user may answer questions such as which genes are up-regulated in normal tissue and down-regulated in diseased tissue. The depicted Entrez definition column contains the definition column from the public domain Entrez database. The depicted query marked "like "growth"" retains those records having the string "growth" as a substring in the designated column.

One condition satisfying the depicted query is that a gene have an expression level in experiment 4002736D greater than 10 and an expression level in experiment 4003228A greater than 10 and less than 0.6 times the expression level in experiment 4002736D. An alternate condition satisfying the query is that the expression level in experiment 4002736D be greater than 10 and the expression level in experiment 4003228A greater than 10 and greater than 1.4 times the expression level in experiment 4002736D.

This query determines the genes that have a particular fold change pattern between experiment 4003228A and experiment 4002736D. It will filter out genes for which there is no significant fold change between the experiments. Specifically, it finds all genes for which the expression level of experiment 4003228A is less than 60% of the expression level of experiment 4002736D, or for which the expression level of experiment 4003228A is greater than 140% of the expression level of experiment 4002736D. Both experiments are also constrained to have expression levels greater than 10.

Filters may be saved or reset by selection of buttons 560 and 562, respectively. The records displayed in lower table 534 may be sorted on any column(s), and columns may be hidden, frozen, or repositioned for better viewing. Lower table 534 may also be saved in different formats, including a spreadsheet format such as Microsoft Excel, by clicking on an export button 564. A saved filter may be accessed via a pull down menu 566 or deleted by selection of a delete button 568. Additional information on any gene may be obtained by double clicking its row. This will load an Internet browser program and open a web site such as the Entrez web site that stores information for the gene. The browser program then displays the entry for that gene.

At step 412, by selecting a graph button 570, the user calls up a scatter-plot display 572 depicted in Fig. 5I. Two experiments are selected for comparison using drop-down lists 574 and 576 for the x axis and y axis respectively. The graph is generated by selecting a build scatter button 578. Each point on the scatter plot corresponds to a particular gene. The point is positioned on the graph according to its measured expression level in both experiments. By checking a box 580, the user may select to have the points color coded according to whether the gene was present in both (2P), one (1P), or neither (OP) of the experiments. By checking one or more of boxes 582, the user may elect to show or not show genes according to this categorization.

10 By making an appropriate selection in a box 584, the user may select an interpretation for future mouse clicks. One choice is for the system to do nothing in response to a mouse click. Another choice is for the system to show gene data for a point selected by a mouse click. The gene data appears in a box 586 including the accession number, the gene name, the expression levels as measured in a variety of experiments, and an expression call.  
15 for each experiment (either absent or present.) An Entrez definition name is also shown. Double clicking on an entry will invoke an Internet browser to show the Entrez entry for the gene.

The user may also select "rope" in box 584 to collect interesting points for comparison by surrounding them with a polygon. Lines are automatically drawn between  
20 each mouse click, encircling those genes to be included in a bar graph. The user may display the bar graph by selecting a button 588.

At step 414, Fig. 5J depicts a bar graph 590 for the roped genes in the scatter plot of Fig. 5I. Each grouping of bars in Fig. 5J corresponds to a gene. Each bar within a grouping corresponds to an experiment and is color-coded according to a legend 592.  
25 Initially only two experiments are displayed, the two experiments corresponding to the axes of the scatter plot of Fig. 5I. However, the user may select further experiments from a box 594. Once the desired experiments are selected, the user selects a build button 596 to display the desired bar graph. A table 598 shows the expression levels for the depicted genes.

For the display of Fig. 5J, the option "gene" is selected in a box 600. To view  
30 individual plots of the expression level for each gene as they vary over the experiments, the user may select option "experiment" in box 600 before selecting build button 596. This

produces a line graph 602 as shown in Fig. 5K. The experiments are arranged along the horizontal axis in the order specified in box 594. Each gene has its own trace corresponding to its expression level as it varies over the experiments. A legend 604 identifies the trace for each gene. To change the position of an experiment along the horizontal axis, the user uses  
5 up and down arrows 606 and 608 to change its position. This feature makes it possible to reorder the experiments to reflect additional sequencing knowledge. For example, if the experiments represent a time course such as progression of a disease or treatment, they can be graphically ordered in time sequence. The graph then represents the change in expression level as a function of time for the selected gene. A slider icon 612 allows the user to scroll  
10 along the horizontal axis if line graph 602 does not fit on the screen. A maker check box 614 shows a horizontal line across line graph 602 defining a particular expression level. This allows the user to easily view data points above the selected level.

More information about a gene may be obtained by clicking on any bar in the group. All of the information for the gene will be displayed in a separate window 610 as  
15 shown in Fig. 5L.

In the foregoing specification, the invention has been described with reference to specific exemplary embodiments thereof. It will, however, be evident that various modifications and changes may be made thereunto without departing from the broader spirit and scope of the invention as set forth in the appended claims and their full scope of equivalents. For example, it will be understood that wherever "expression level" is referred to, one may substitute the measured concentration of any compound. Also, wherever "gene" is referred to, one may substitute the term "expressed sequence tag."



WHAT IS CLAIMED IS:

1           1.     In a computer system, a method for operating a database storing  
2 expression level information comprising:  
3           providing a database comprising expression levels for each of a plurality of  
4 genes or expressed sequence tags (EST) as measured in each of a plurality of tissue types;  
5           accepting a user query to said database to identify desired ones of said  
6 plurality of genes or EST, said user query specifying expression level characteristics of said  
7 desired genes; and  
8           comparing said expression level characteristics to said expression levels stored  
9 in said database to identify said desired genes or EST.

1           2.     The method of claim 1 further comprising:  
2 displaying information identifying said desired genes or EST.

1           3.     The method of claim 1 wherein said plurality of tissue types comprise  
2 a diseased tissue type.

1           4.     The method of claim 1 wherein said plurality of tissue types comprise  
2 a healthy tissue type.

1           5.     The method of claim 1 wherein said plurality of tissue types comprise  
2 a cancerous tissue type.

1           6.     The method of claim 1 wherein said plurality of tissue types comprise  
2 a drug treated tissue type.

1           7.     The method of claim 1 wherein said plurality of tissue types comprise  
2 issues obtained from disparate species.

1           8.     The method of claim 1 wherein said plurality of tissue types comprise  
2 tissues obtained from disparate organs.

1           9.     The method of claim 1 wherein said expression level characteristics  
2 comprise expression level ranges as measured for a particular gene in at least two of said  
3 plurality of tissue types.

1           10.    The method of claim 1 wherein said expression level characteristics  
2 comprise relationships among expression levels as measured for a particular gene in at least  
3 two of said plurality of tissue types.

1           11.    The method of claim 1 further comprising:  
2 accepting user input selecting two of said plurality tissue types for graphical  
3 display;  
4 displaying a first axis corresponding to a first one of said two tissue types;  
5 displaying a second axis corresponding to a second one of said two tissue  
6 types;  
7 for a selected one of said plurality of genes or EST, displaying a mark at a  
8 position wherein said position is selected relative to said first axis in accordance with an  
9 expression level of said selected gene or EST measured in said first tissue type and selected  
10 relative to said second axis in accordance with an expression level of said selected gene or  
11 EST measured in said second tissue type.

1           12.    The method of claim 11 further comprising:  
2 repeating said operation of displaying a mark for a plurality of selected genes  
3 or EST.

1           13.    In a computer system, a method for operating a database storing  
2 information about compound concentration comprising:  
3 providing a database comprising concentrations of a plurality of compounds as  
4 measured in a plurality of samples;  
5 accepting a user query to said database to identify desired ones of said  
6 plurality of compounds, said user query specifying concentration characteristics of said

7 desired compounds in selected ones of said plurality of samples; and  
8 comparing said concentration characteristics to said concentrations stored in  
9 said database to identify said desired compounds.

1 14. A computer program product for operating a database storing  
2 expression level information comprising:  
3 code that provides a database comprising expression levels for each of a  
4 plurality of genes or expressed sequence tags (EST) as measured in each of a plurality of  
5 tissue types;  
6 code that accepts a user query to said database to identify desired ones of said  
7 plurality of genes or EST, said user query specifying expression level characteristics of said  
8 desired genes;  
9 code that compares said expression level characteristics to said expression  
10 levels stored in said database to identify said desired genes or EST; and  
11 a computer-readable storage medium for storing the codes.

1 15. The product of claim 14 further comprising:  
2 code that displays information identifying said desired genes or EST.

1 16. The product of claim 14 wherein said plurality of tissue types comprise  
2 a diseased tissue type.

1 17. The product of claim 14 wherein said plurality of tissue types comprise  
2 a healthy tissue type.

1 18. The product of claim 14 wherein said plurality of tissue types comprise  
2 a cancerous tissue type.

1 19. The product of claim 14 wherein said plurality of tissue types comprise  
2 a drug treated tissue type.

1           20.    The product of claim 14 wherein said plurality of tissue types comprise  
2 tissues obtained from disparate species.

1           21.    The product of claim 14 wherein said plurality of tissue types comprise  
2 tissues obtained from disparate organs.

1           22.    The product of claim 14 wherein said expression level characteristics  
2 comprise expression level ranges as measured for a particular gene in at least two of said  
3 plurality of tissue types.

1           23.    The product of claim 14 wherein said expression level characteristics  
2 comprise relationships among expression levels as measured for a particular gene in at least  
3 two of said plurality of tissue types.

1           24.    The product of claim 14 further comprising:  
2           code that accepts user input selecting two of said plurality tissue types for  
3 graphical display;  
4           code that displays a first axis corresponding to a first one of said two tissue  
5 types;  
6           code that displays a second axis corresponding to a second one of said two  
7 tissue types;  
8           code that, for a selected one of said plurality of genes or EST, displays a mark  
9 at a position wherein said position is selected relative to said first axis in accordance with an  
10 expression level of said selected gene or EST measured in said first tissue type and selected  
11 relative to said second axis in accordance with an expression level of said selected gene or  
12 EST measured in said second tissue type.

1           25.    The product of claim 24 further comprising:  
2           code that repeatedly applies said code that displays a mark for a plurality of  
3 selected genes or EST.

1           26.    A computer program product for operating a database storing  
2 information about compound concentration comprising:  
3           code that receives a database comprising concentrations of a plurality of  
4 compounds as measured in a plurality of samples;  
5           code that accepts a user query to said database to identify desired ones of said  
6 plurality of compounds, said user query specifying concentration characteristics of said  
7 desired compounds in selected ones of said plurality of samples; and  
8           code that compares said concentration characteristics to said concentrations  
9 stored in said database to identify said desired compounds.

1           27.    A computer system comprising:  
2 a processor; and  
3 a memory storing code to operate said processor, said code comprising:  
4 code that provides a database comprising expression levels for each of a  
5 plurality of genes or expressed sequence tags (EST) as measured in each of a plurality of  
6 tissue types;  
7           code that accepts a user query to said database to identify desired ones of said  
8 plurality of genes or EST, said user query specifying expression level characteristics of said  
9 desired genes; and  
10           code that compares said expression level characteristics to said expression  
11 levels stored in said database to identify said desired genes or EST.

1 / 15

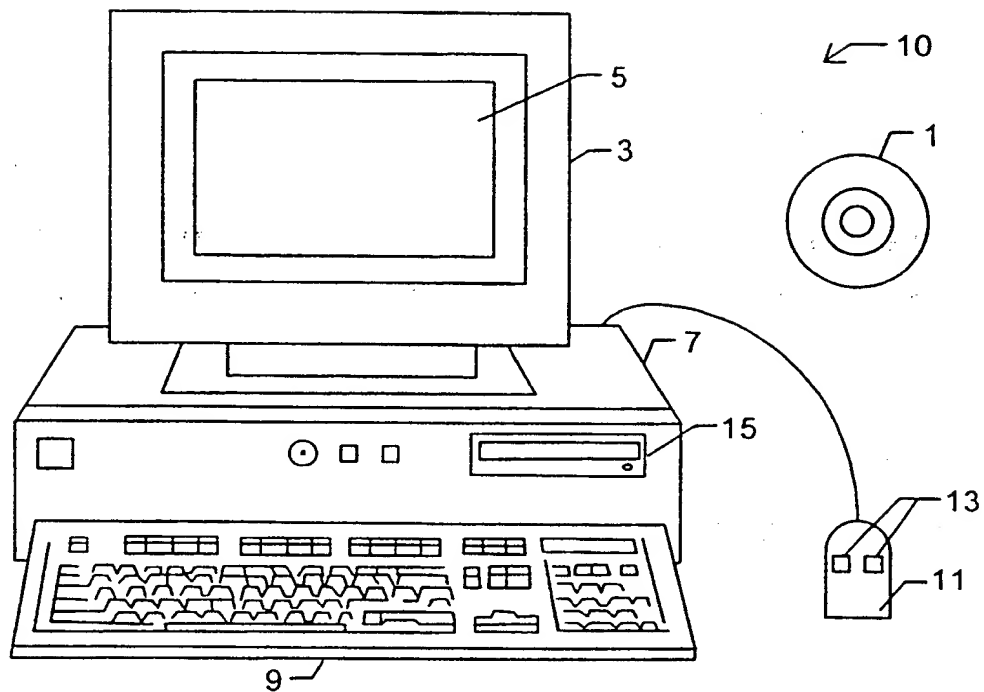


FIG. 1

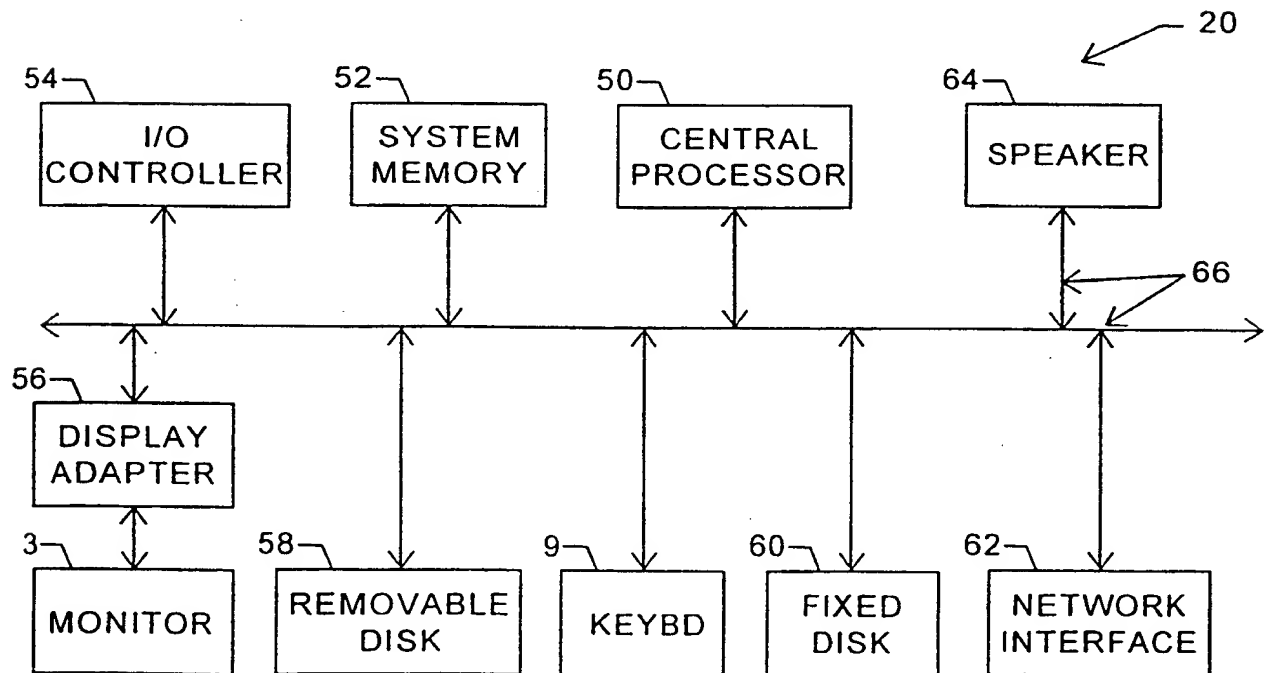


FIG. 2

2 / 15

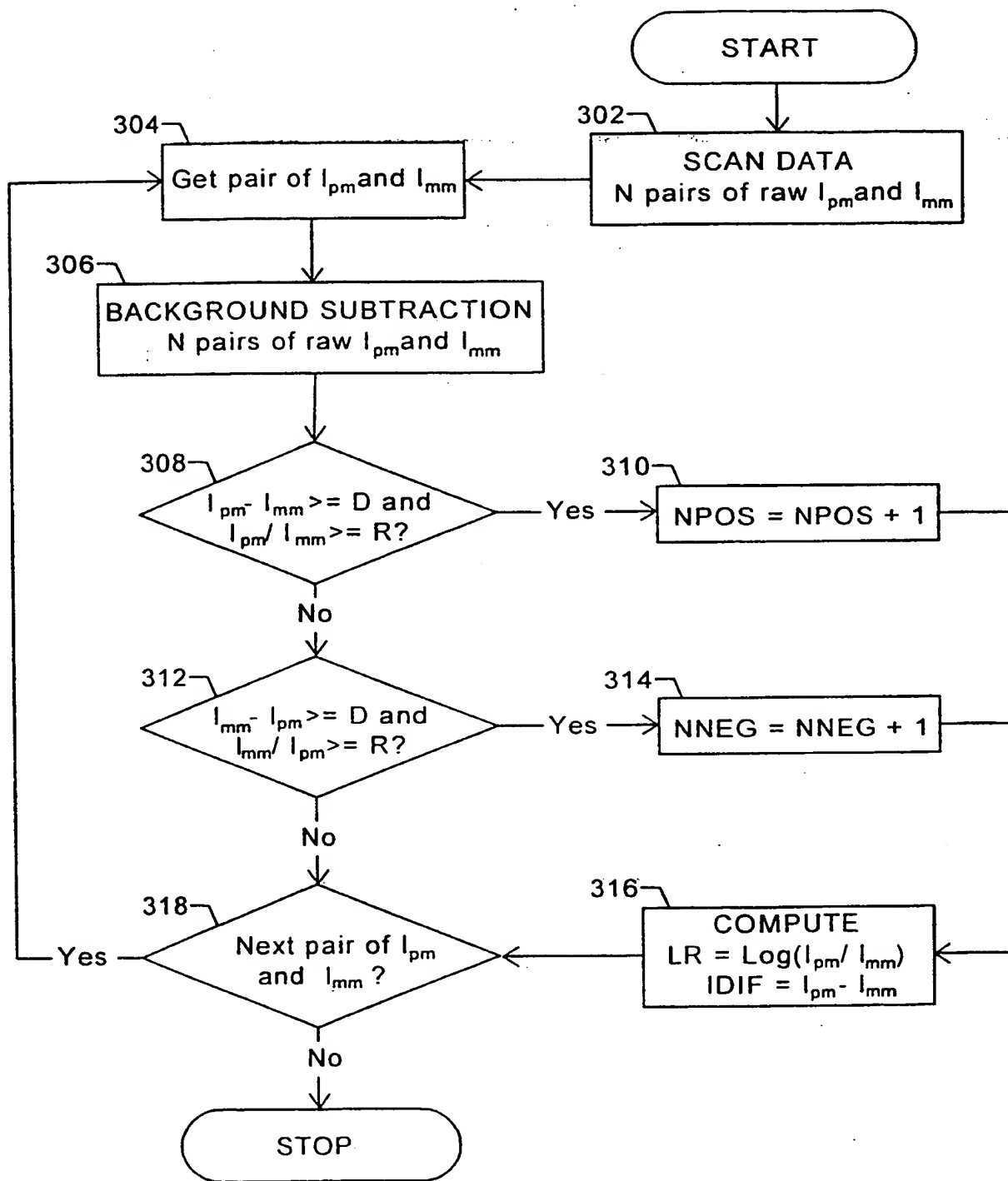


FIG. 3

3 / 15

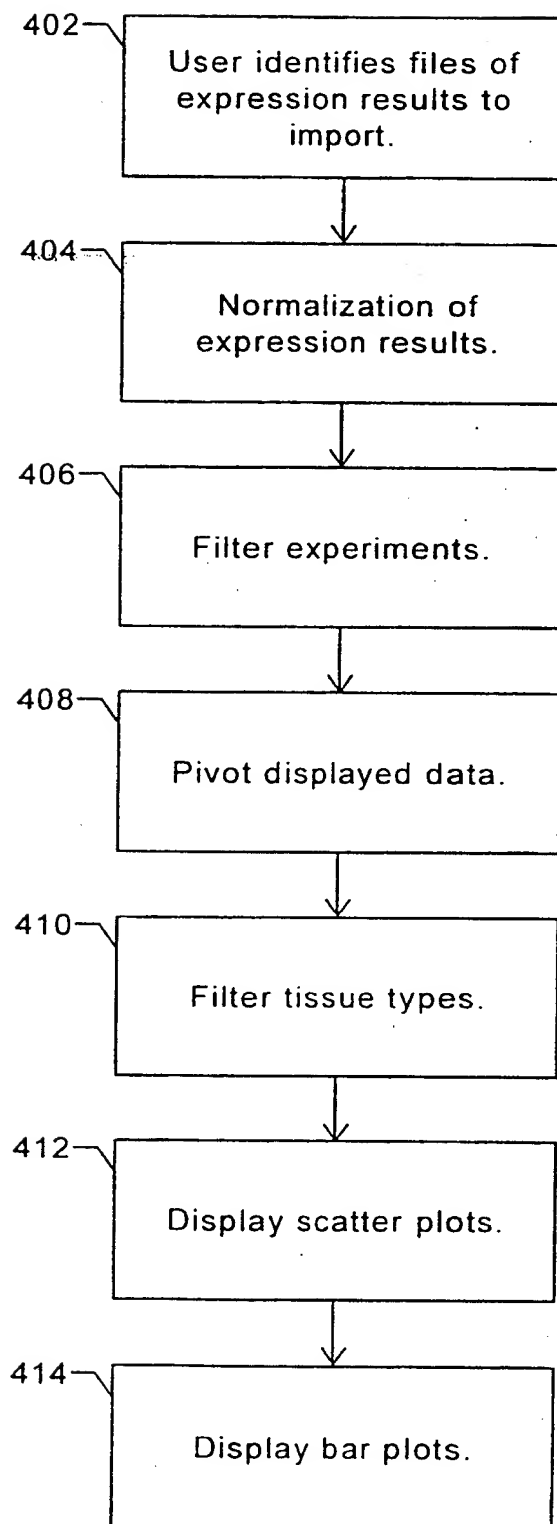


FIG. 4



504

506

514

502

512

508

Import and normalization

Import

Normalization

select files

Analysis Type

Species

File List

delete file list

save file list

import

reset file list

Gene Chip Results

FileName	Code	Replicate	ChipDesignCode	Sam
g:\genexel\Gish Data\text files from CHP files\y050103.txt	4002736D	0	A	cand
g:\genexel\Gish Data\text files from CHP files\y050201.txt	4002736D	0	B	cand
g:\genexel\Gish Data\text files from CHP files\y050601.txt	4002736D	0	C	cand
g:\genexel\Gish Data\text files from CHP files\y050701.txt	4002736D	0	D	cand
g:\genexel\Gish Data\text files from CHP files\y080501.txt	4003228A	0	A	cand
g:\genexel\Gish Data\text files from CHP files\y080601.txt	4003228A	0	B	cand
g:\genexel\Gish Data\text files from CHP files\y080701.txt	4003228A	0	C	cand
g:\genexel\Gish Data\text files from CHP files\y080801.txt	4003228A	0	D	cand
g:\genexel\Gish Data\text files from CHP files\y052001.txt	4003265J	0	A	cand

Gene Expression Levels

FIG. 5A

**Import and normalization**

**Import**

**select files** ▼

Analysis Type ▼

Species ▼

File List ▼

Human

Kurt's Famous 48

**delete file list** import

**save file list** reset file list

**Normalization**

**Gene Chip Results**

FileName Code Replicate ChipDesignCode Serial

516

Select this file

Look in: text files from CHP files

File Name	File Type	File Size	File Date	File Time	File Owner	File Group	File Permissions	File Attributes	File Comments
sy040901.txt	sy042201.txt	sy043003.txt	sy042403.txt	sy042404.txt	sy042501.txt	sy042502.txt	sy042503.txt	sy042504.txt	sy042901.txt
sy040902.txt	sy042202.txt	sy043004.txt	sy042404.txt	sy042501.txt	sy042502.txt	sy042503.txt	sy042504.txt	sy042901.txt	sy042902.txt
sy040904.txt	sy042203.txt	sy050101.txt	sy050103.txt	sy050104.txt	sy050201.txt	sy050202.txt	sy050205.txt	sy050601.txt	sy050602.txt
sy041501.txt	sy042204.txt	sy050103.txt	sy050104.txt	sy050201.txt	sy050202.txt	sy050205.txt	sy050601.txt	sy050602.txt	
sy041502.txt	sy042301.txt	sy050103.txt	sy050104.txt	sy050201.txt	sy050202.txt	sy050205.txt	sy050601.txt	sy050602.txt	
sy041503.txt	sy042302.txt	sy050103.txt	sy050104.txt	sy050201.txt	sy050202.txt	sy050205.txt	sy050601.txt	sy050602.txt	
sy041504.txt	sy042303.txt	sy050103.txt	sy050104.txt	sy050201.txt	sy050202.txt	sy050205.txt	sy050601.txt	sy050602.txt	
sy041704.txt	sy042304.txt	sy050103.txt	sy050104.txt	sy050201.txt	sy050202.txt	sy050205.txt	sy050601.txt	sy050602.txt	
sy041801.txt	sy042401.txt	sy050103.txt	sy050104.txt	sy050201.txt	sy050202.txt	sy050205.txt	sy050601.txt	sy050602.txt	
sy041802.txt	sy042402.txt	sy050103.txt	sy050104.txt	sy050201.txt	sy050202.txt	sy050205.txt	sy050601.txt	sy050602.txt	

Select Cancel Advanced...

Find files that match these criteria:

File name: Text or property: ▼

Files of type: Last modified: ▼

any time

96 file(s) found.

Find Now New Search

ression Levels

FIG. 5B

518

519

# Import and normalization

Import

Normalization

select files

Analysis Type

Species

File List

delete file list

import

save file list

reset file list

Gene Chip Results

FileName	ChipDesignCode	Sample
g:\genexel\Gish Data\text files from CH	0 C	cand
g:\genexel\Gish Data\text files from CH	0 A	cand
g:\genexel\Gish Data\text files from CH	0 B	cand
g:\genexel\Gish Data\text files from CH	0 D	cand
g:\genexel\Gish Data\text files from CH	0 C	cand
g:\genexel\Gish Data\text files from CH	0 A	cand
g:\genexel\Gish Data\text files from CH	0 B	cand
g:\genexel\Gish Data\text files from CH	0 D	cand
g:\genexel\Gish Data\text files from CH	0 C	cand

## Gene Expression Levels

FIG. 5C

Import and normalization
Gene Expression Levels

Import
Normalization

Chip Variability

	Gene Name	Accession #
<input type="checkbox"/>	HUMGAPDH/M33197	HUMGAPDH/M33197
<input type="checkbox"/>	HSAC07/X00351	HSAC07/X00351
<input checked="" type="checkbox"/>	▲	*

Scale Value

☐ Use
☒ Don't Use

apply

Tissue Variability

Scale Value

☐ Use
☒ Don't Use

Transform

Negative to log

☒ Use
☐ Don't Use

reset

FIG. 5D

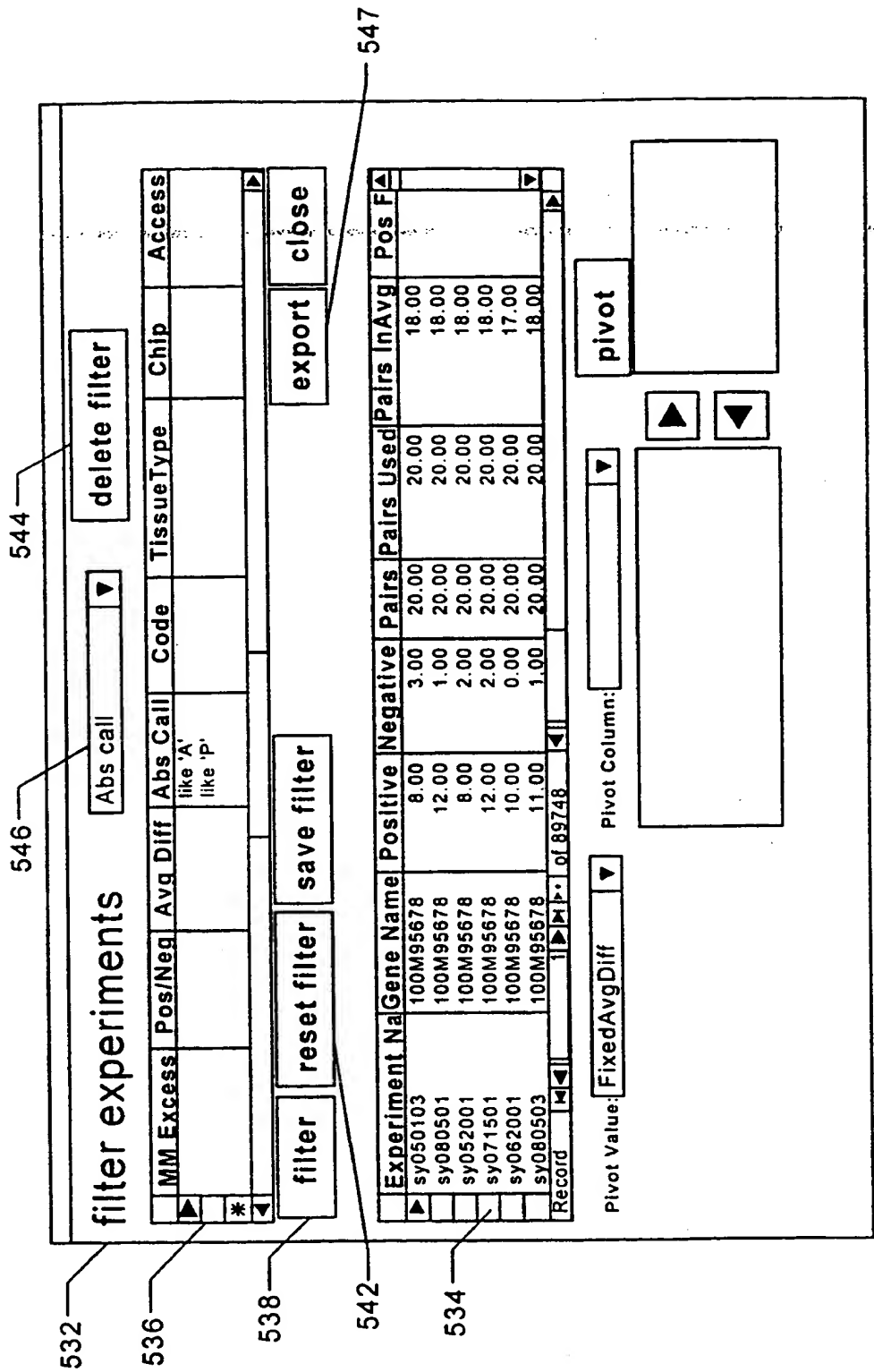


FIG. 5E

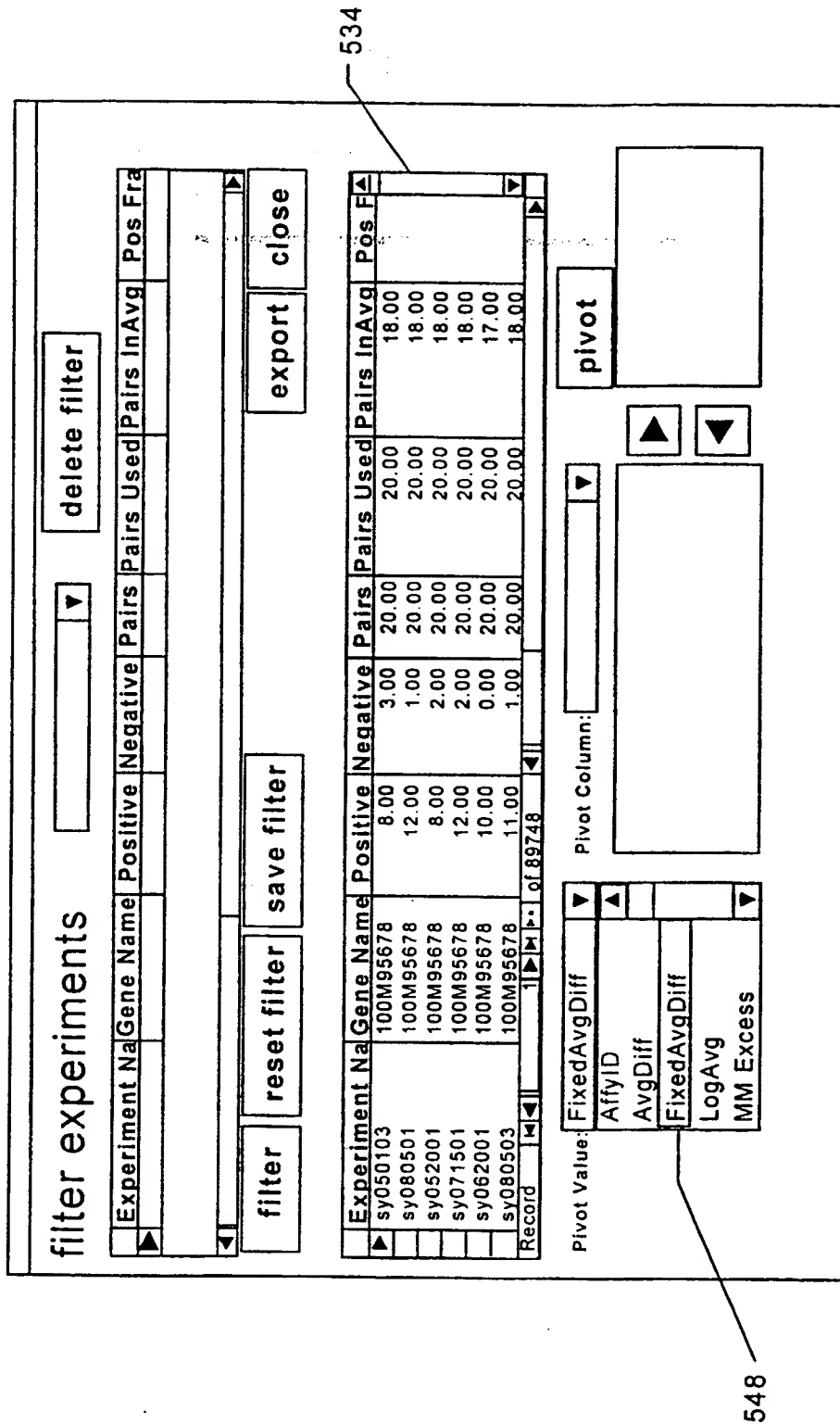


FIG. 5F

filter experiments

Experiment Na	Gene Name	Positive	Negative	Pairs Used	Pairs InAvg	Pos Fra
sy050103	100M95678	8.00	3.00	20.00	18.00	
sy080501	100M95678	12.00	1.00	20.00	18.00	
sy052001	100M95678	8.00	2.00	20.00	18.00	
sy071501	100M95678	12.00	2.00	20.00	18.00	
sy062001	100M95678	10.00	0.00	20.00	17.00	
sy080503	100M95678	11.00	1.00	20.00	18.00	

Record 1 of 89748

Pivot Value: FixedAvgDiff   
Pivot Column: Code

pivot
4002736D
4003228A
4003265J
4003343E
4003344H

FIG. 5G

filter tissue types

renew with growth

delete filter

Gene Name	4002736D	4003228A	Entrez Definition	4003265J	4003343E	400334
	>10	>10 and <0.6*[4002736D]like "growth"				
	>10	>10 and >1.4*[4002736D]like "growth"				

filter

reset filter

save filter

graph

export

close

538

562

536

564

534

Gene Name	4002736D	4003228A	4003265J	4003343E	4003344H	4003353C	400334
1000M95678	424.14	371.34	220.06	301.90	258.31	379.82	
1000T51617	10.59	15.85	0.06	19.48	33.87	31.72	
1001M83363	10.36	55.68	0.03	34.66	34.81	57.97	
1002R06239	974.83	141.27	62.31	80.49	103.63	88.36	
1003L14922	1.23	27.92	11.11	1.69	0.08	0.09	
1003M32402	0.04	0.06	0.06	0.08	0.07	0.05	
1004T51944	61.38	18.65	15.25	15.36	22.07	5.83	
1004L02932	0.03	13.86	0.30	6.74	0.04	1.19	
1004T51849_f	350.21	164.89	134.93	237.75	137.96	334.93	
1004T51849_i	297.03	0.02	0.01	84.67	11.77	0.00	
10045T51849_r_i	224.27	0.01	0.01	81.43	0.01	0.00	
1005L07592	71.98	119.08	86.34	118.06	97.49	167.91	

Record

1

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

FIG. 5H



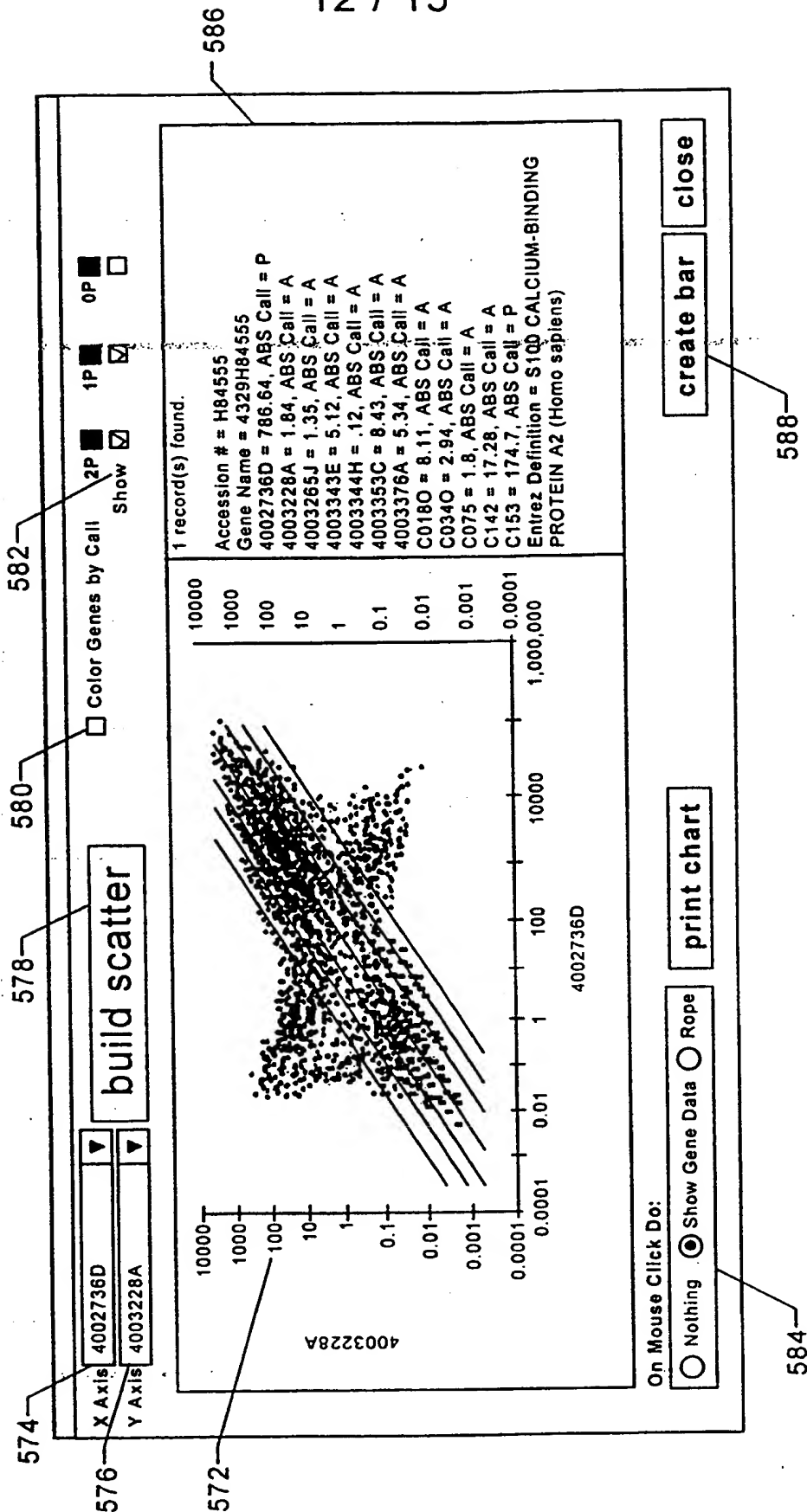


FIG. 5I

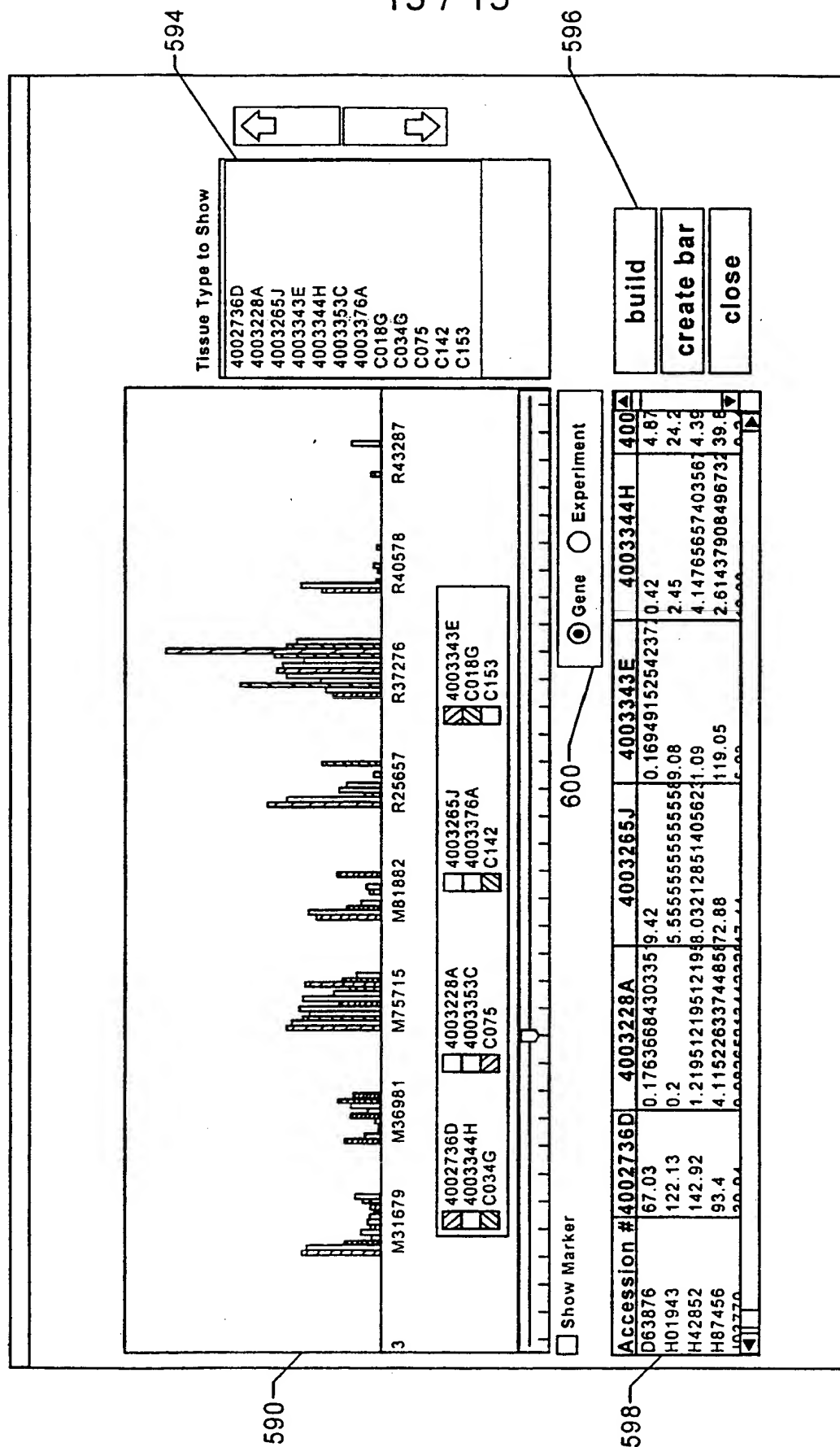


FIG. 5J

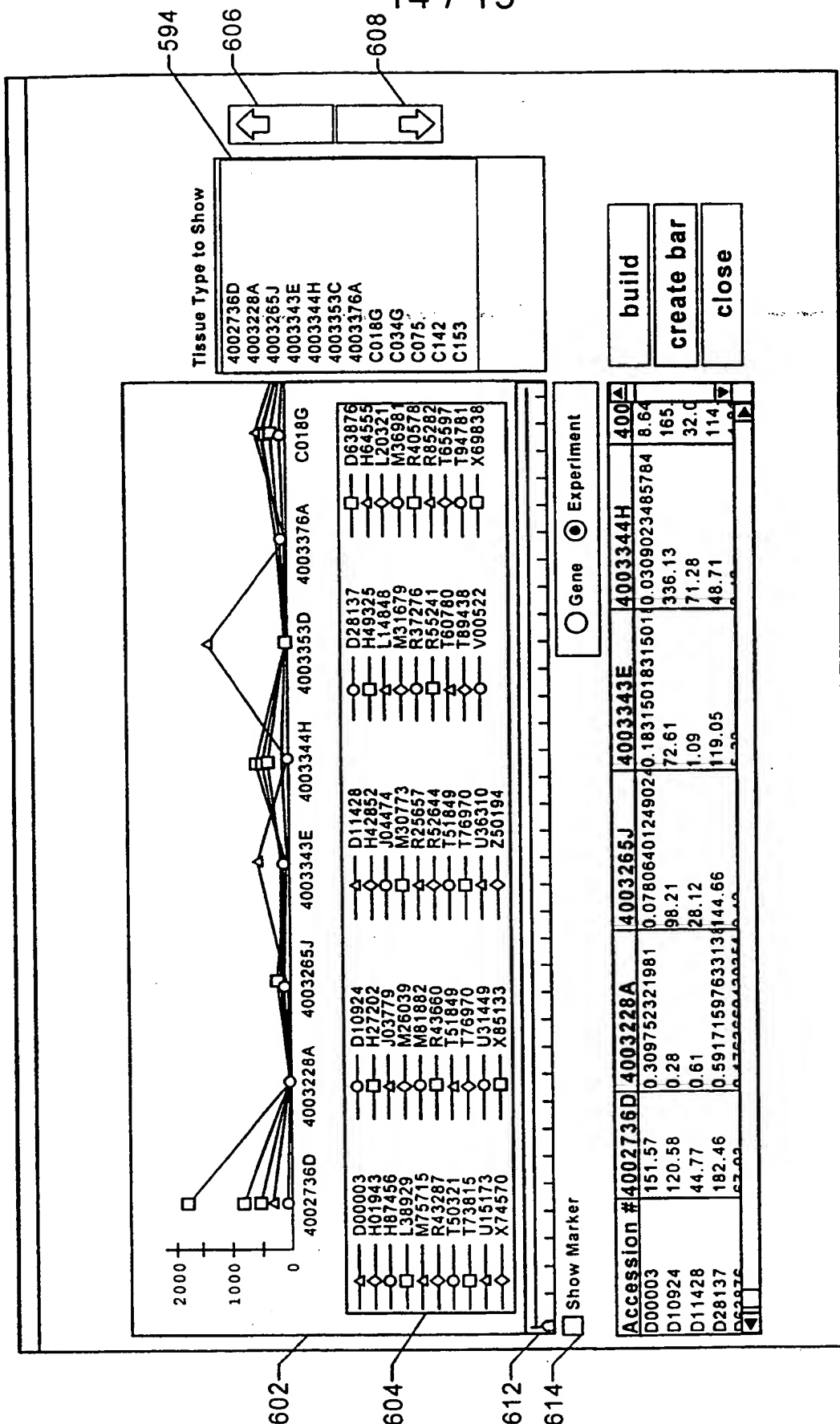


FIG. 5K

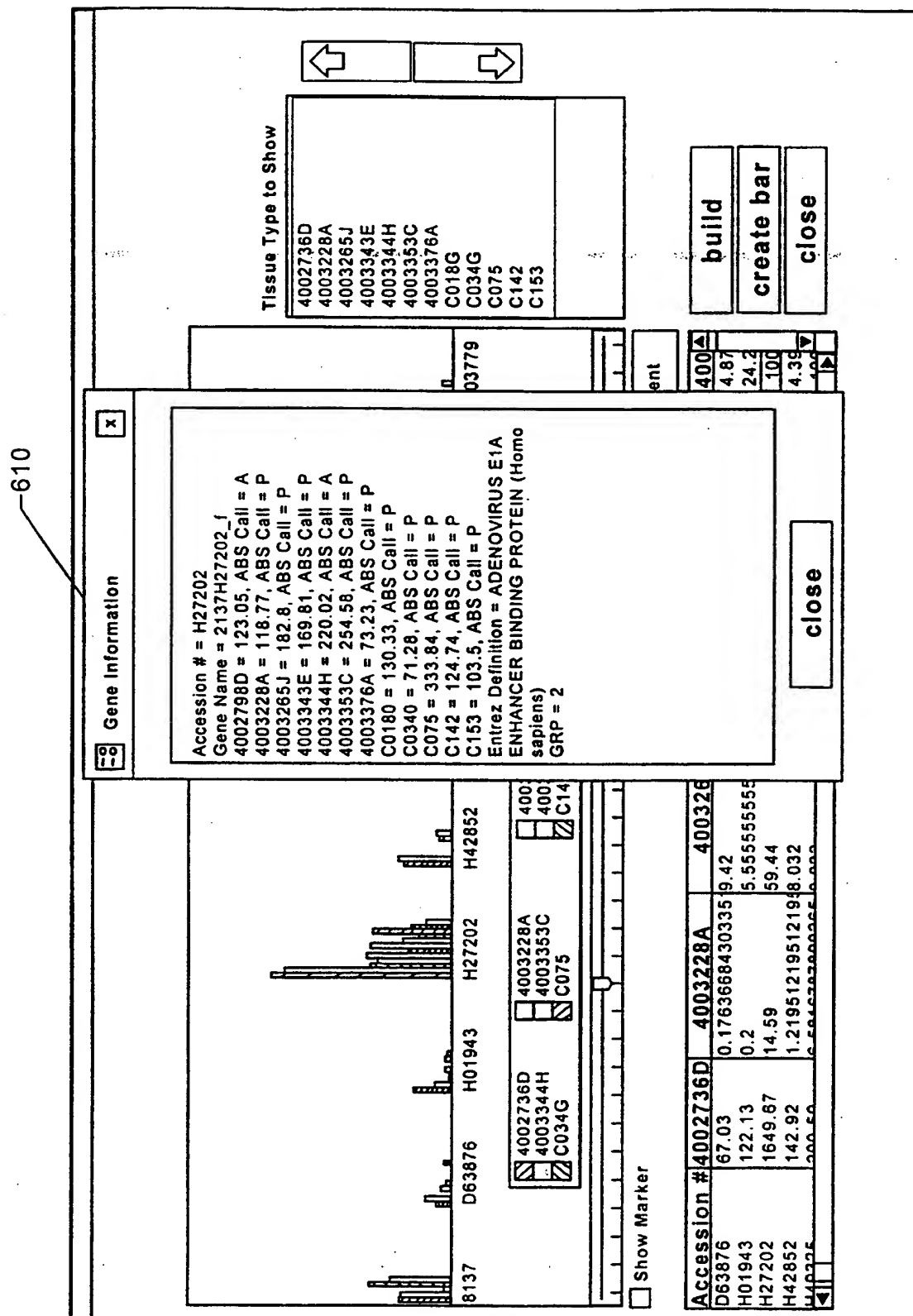


FIG. 5L

## INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US98/15151

## A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68; G06F 17/30, 159/00

US CL :435/6; 707/1; 128/923

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 707/1; 128/923

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, STN. Search terms: transcript, transcripts, cdna#, mrna#, est#, frequenc?, abundanc?, distribut?, computer#, algorithm#

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X - Y	OKUBO et al. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression, Nature Genetics. November 1993, Vol. 2, No. 3, pages 173-179, see entire document.	1, 2, 4, and 8-10 <u>3</u> , 5-7, and 11-27
Y	IntelliGenetics Suite (TM), Release 5.4, Advanced Training Manual, January 1993, published by IntelliGenetics, Inc., 700 East El Camino Real, Mountain View, California 94040, USA, pages (1-6) - (1-19) and (2-9) - (2-14), see entire document.	3, 5-7, and 11-27
X - Y	WO 96/23078 A1 (INCYTE PHARMACEUTICALS, INC.) 01 August 1996, see entire document.	1-10 and 13-27 <u>11</u> and 12



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:	*T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
*A* document defining the general state of the art which is not considered to be of particular relevance	*X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
*B* earlier document published on or after the international filing date	*Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
*L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*A* document member of the same patent family
*O* document referring to an oral disclosure, use, exhibition or other means	
*P* document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

16 OCTOBER 1998

Date of mailing of the international search report

29 OCT 1998

Name and mailing address of the ISA/US  
Commissioner of Patents and Trademarks  
Box PCT  
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

JAMES MARTINELL *for*

Telephone No. (703) 308-0196

